User-Defined Gestural Interaction: a Study on Gesture Memorization

Jean-François Jégo, Alexis Paljic, Philippe Fuchs

User-Defined Gestural Interaction: a Study on Gesture Memorization

Jean-François Jégo[1]
Mines ParisTech, Paris, France

Alexis Paljic[2]
Mines ParisTech, Paris, France

Philippe Fuchs[3]
Mines ParisTech, Paris, France

ABSTRACT

In this paper we study the memorization of user created gestures for 3DUI. Wide public applications mostly use standardized gestures for interactions with simple contents. This work is motivated by two application cases for which a standardized approach is not possible and thus user specific or dedicated interfaces are needed. The first one is applications for people with limited sensory-motor abilities for whom generic interaction methods may not be adapted. The second one is creative arts applications, for which gesture freedom is part of the creative process. In this work, users are asked to create gestures for a set of tasks, in a specific phase, prior to using the system. We propose a user study to explore the question of gesture memorization. Gestures are recorded and recognized with a Hidden Markov Model. Results show that it seems difficult to recall more than two abstract gestures. Affordances strongly improve memorization whereas the use of colocalization has no significant effect.

Keywords: User studies, 3D interaction, gesture, usability.

Index Terms: I.3.6 [Methodology and Techniques]: Interaction techniques; I.3.7 [Three-Dimensional Graphics and Realism]: Virtual reality

1 GENERAL CONTEXT

The context of this work is user-defined gestures for interaction with digital content, in Virtual Reality (VR) or desktop setups. By user-defined gestures, we point out interactive systems in which the user is free to propose the gestures for a given set of tasks. In the literature, there are three gestural interaction approaches: standardized, adapted [1] and user-derived [2].

An example of standardized gestural interaction can be seen in wide public applications, such as gaming consoles, information kiosks or television sets [3] that propose gestural control. For the standardized interaction, the system imposes to the user the set of gestures and thus makes him go through a learning process. In the adapted approach the user can tune to some extent the interactions to match his preferences [4].

However, there are specific use cases in VR or 3DUI where these two bottom-up approaches (standardized, adapted) are not possible. The first is for persons with motor disabilities, for whom the standard interactive gestures may not be suitable. For example, there is a specific need for user-defined interfaces for patient rehabilitation applications using VR. Secondly, creative applications, in which artists can create their own gestural commands to control media in performing arts.

Our idea is that a top-down approach, from the user towards the system, is needed to really adapt to user needs in these two applications cases. In this paper, we want to explore a user-defined interaction also called user-derived interaction [2] that allows the user to make the system learn the gestures that he has created, in a specific phase, prior to using the system.

This raises the question of the number of different gestures that the user can create and memorize. Literature on working memory suggests that only a limited number of items can be memorized. However, there are few works on gesture memorization for interactive applications. In order to perform a multitude of actions in a Virtual Environment (VE), one needs to know whether there is a limit in the number of gestures that a user can create and reuse and what can influence the ability to memorize them.

Thus, we want to explore what parameters facilitate the recall of user-defined gestures. It is our hypothesis that the semantic content of the gesture and its relation to the object or concept at hand affects gesture memorization. This introduces the concept of schema (plural schemata), which is the innate knowledge of how to use a specific object, based on experience. In this work, we point out two important characteristics of schema: affordances and colocalization. Affordances are characteristics of objects that make their use self-explanatory. Colocalization in VR is the fact that visualization space and manipulation space are superimposed.

We propose a user study on user-defined gestures memorization. The user study is based on a task consisting in opening various virtual boxes. Our platform allows the user to create a gesture for each box.

In the Previous Work & Definitions section of this paper, we expose the main approaches for gestural interaction. We also present works on memorization that can have an influence on such an approach and finally we define schema and its characteristics. In the User Study section, we present our hypothesis, the setup and the protocol. Gesture memorization results are then presented and discussed in the light of two parameters: 1) the presence or not of affordances in the VE and 2) co-localized versus indirect manipulation. An Appendix at the end of this paper presents technical information on the gesture following algorithms used in the experiment.

2 PREVIOUS WORK & DEFINITIONS

2.1 Standardized interaction

In the field of VR, Cabral et al. [5] have proposed a standardized 2D set of gestures for application control, selection and manipulation of objects in a large scale 3D environment. Authors report that the large amplitude of proposed movement causes fatigue, but also that the interface is easy to use and learn and is appropriate for short sporadic use. Bobiller-Chaumon et al. [1], noticed that standardized interaction benefits from a great accessibility due to common cultural references and from a possible transfer of competences. It can be reused between applications and is also easier to implement. In the past years, some information kiosks in museums or shopping malls have proposed gestural control. Most gaming consoles have gestural control [6] that allows natural movements in sports or serious

[1]e-mail: jean-francois.jego@mines-paristech.fr
[2]e-mail: alexis.paljic@mines-paristech.fr
[3]e-mail: philippe.fuchs@mines-paristech.fr

games. Another example of this trend is the use of gestural control for consumer electronics such as media/menu navigation in TVs [3] through *innate* gestures, like pointing, grabbing or learned ones like swiping or zooming. Users exposed to these devices tend to acquire experience and reuse it when they face a new device or application in a *propagation effect*. However, Norman [7] noticed that accessibility is limited because standardized interaction is designed for a generic purpose and cannot take into account inter-individual or cultural differences. For example waving the hand to engage on a game console means "hello" in western cultures but it means rejection, disapproval or lack of interests for Indian users.

One can think that these systems provide the user with the freedom of choosing the movements which seem natural to him. In fact, the movements are strongly implied by the context in the VE, with strong feedback, helpers or visual clues as pictograms, animations, tutorials, etc. [8]. Also, these devices have a strong gesture error tolerance, within a general system-defined gesture pattern. The constraint here is for the system to be adapted to a large public and, for the same reasons, designed for a reduced set of tasks.

Thus, it seems interesting to have an interaction process that lets the user customize parameters to interact as naturally as possible, depending on his habits and system expectations. To do so, a possible solution is to adapt interaction.

2.2 Adapted interaction

The continuum proposed by Oppermann [9] describes a continuous scale between adaptive and adaptable interactions. Adaptive or dynamic systems adapt to the users automatically based on the assumptions of the system about user needs, whereas adaptable or static systems allow the user to change certain system parameters (the median approach lets the user select the adaptation suggested by the system). A common desktop example of an adaptable system is the customization of 2D interactions such as changing the color, the size or the speed of a cursor. In the field of VR, Bowman et al. talk about adding, modifying or tweaking techniques to produce flavors [4]. Adaptation can increase comfort and efficiency during interaction since it fits better the user specificities. Bobillier-Chaumon et al. [1] point out the advantages of flexibility, versatility and context-awareness. However, there are strong risks of isolation (i.e. no reuse for other applications), complexity of implementation and evaluation. They also noticed possible negative effects of an incorrect adaptation.

The choice of adaptation versus standardization is an open question. For a wide public application with a small set of tasks, a standardized approach including helpers to reduce or avoid a learning step seems preferable. For advanced or dedicated needs, a system-predefined set of gestures based on consensus studies, with a user learning step, may be more suitable. However, Oppermann's continuum does not consider the case where the user defines the interaction process himself. The continuum only focuses on who controls the parameters of the predefined interaction, system or user.

2.3 User-defined interaction

In this work, we focus on a third approach that is user-defined interaction also called user-derived interaction [2]. Recent works in VR look towards sensorimotor or cognitive rehabilitations such as stroke rehabilitation [10]. The goal is to design systems assisting the patients in daily activities and autonomy, with the objective of being adapted to patients. Indeed, people with cognitive or motion impairments cannot interact as simply as valid persons. And this turns out to be harder in VEs, because of the high cognitive loads required in the interaction process, including the interface learning step [11]. The second application domain of user-defined gestures is creative arts where a performer's body movement creates content or controls agents. Here, the freedom and multiplicity of movements are required as they are parts of the creative process. There are examples of user-defined controls for artistic creation, such as the Kinectar Performance Platform [12] that allows defining customized virtual instruments and the associated gestures through a functionality called *instrument builder*. An approach where setting the interfaces according to user's abilities or desires and where gestures are proposed to interact, would by definition be more suitable.

Few studies explored the user-defined approach. A close work is a guessability study for mobile phone interaction by Ruiz et al. [13]. The authors show that a consensus exists between users who were asked to define movements to invoke commands on a hand held mobile device. Another study by Wobbrock et al. [14] proposes a user-defined gestures approach to tabletop interaction: they found that desktop idioms strongly influence user's mental models and that some commands elicit little consensus.

3 Memorization of Self-Defined Gestures

3.1 Working memory

A main limit of user-defined gestural interaction is the ability of the user to memorize the gesture set. The theoretical concept of working memory assumes information to be held temporarily in an accessible state. This supports some human mental tasks and provides an interface between perception, long-term memory and action [15]. Memorizing a set of abstract gestures evokes the process of memorizing a list of numbers. Miller's law [16] reveals that the working memory allows remembering only five to nine numbers. Thus, LaViola et al. [8] presume that the number of created gestures that one can remember could be around 7. In fact, recent studies [17] show that working memory is closer to three or four items.

Wagner et al. [18] studied the impact of making gestures in order to remember a string of letters or a visual grid pattern. The study involved a verbal math resolution problem with or without gestures; prior to this a set of items had to be memorized. Participants remembered significantly more items using the gestures than without them. When gesture conveyed the same propositional information as speech, more items were remembered.

3.2 Schema/Schemata

The action performed in the process of using something in daily life activities without learning how to use it is called a schema. The concept was proposed by psychologist Piaget [19]. It refers to the mental organization of actions as they are transferred or generalized while repeating this action in similar circumstances. In his sensorimotor activity, the person uses schemata that he or she acquired, as experiences. As observed by Piaget, schemata are reproducible, have a purpose and are used and assimilated unconsciously. Hence, they do not require a training process for basic activities.

In the field of VR, due to technical or economic reasons, it is not always possible to implement schema for interaction. Fuchs et al. [20] suggest that instead of using schemata, symbolic interaction methods called metaphors can be used instead but have to be learned by the user. These are for instance largely used in standardized interaction modalities. In order to avoid cognitive load, a user-defined approach could focus on the use of schemata.

3.3 Affordances

A way to evoke schemata to the user is to bring to mind clues that provide directions of use for objects or situation. These clues are called affordances by Gibson [21] and refer to the action possibilities provided by an artifact/environment, which may or may not be perceived by the user. Mullaly [22] proposed to provide affordances to desktop software, also called skeuomorphisms, adding a realistic look to elements of the application, e.g. a dialer application that looks like a real dialer. Visual affordances help novice users to interact whereas advanced users may disregard them.

Regarding manipulation tasks, Rizzolatti et al. [23] define them as the elements for constructing possible motor grasps depending on the relative position of the object to our own body. Smets et al. [24] noticed that coupling direct manipulation with affordances is a way to improve the design of objects in VEs. This suggests that colocalization could have a role in the perception of affordances or schemata and thus on gesture memorization.

3.4 Colocalization

Colocalization, also called direct manipulation in the literature, is characteristic of situations where manipulation and visualization spaces are superimposed. The situation where there is an offset between these two spaces is called distant or indirect manipulation. Several studies show that direct manipulation is preferable to distant manipulation.

According to Mine [25], a user can take advantage of proprioception during body-relative interaction in at least three ways: direct manipulation, physical mnemonics and gestural actions. Also, it is needed to provide interactions within arm's reach, i.e. within a user's natural working volume. This provides a more direct mapping between hand motion and object motion and a finer angular precision of motion.

In the case of pointing tasks, Paljic et al. [26] showed that direct and close manipulations are more efficient than indirect manipulations (distance > 40 cm). 3D cursor speed and visual clues, such as a ray that indicates manipulation offset, also decreased the performance. This suggests that the environment and objects should avoid artificial or accessory clues, especially if we focus on schemata.

Previous work shows that interaction can be proposed by the system and adapted to the user; on the other side, the user can define his own interaction modalities. For advanced-usage or to fit user's possibilities, gestures will thus be acquired in the range of his own sensory-motor and cognitive abilities. The question of variety of gestures and their memorization is central to this work. Our hypothesis is that if the users propose spontaneous gestures that are inspired by everyday life gestures, it may enhance memorization.

4 User Study

The user-defined gestural interaction method that we propose consists in two specific phases. The first one is **gesture creation** and system learning. The second phase is **gesture interaction** which is the actual use of these gestures within the application.

Our hypothesis is that the gesture memorization is facilitated using schemata, i.e. with explicit affordance on object and direct manipulation. We also hypothesize that an indirect and non affordant manipulation would elicit a poorer gesture memorization.

4.1 Proposed task

We chose to apply the user-defined approach to an ecological bimanual box opening task in a VE (Figure 1). The boxes are presented with or without affordances and within or without arms reach (Figure 2). In this experiment, we chose not to show box opening animations. The reason of this choice is to focus first on static visual affordances related to the manipulated object.

Since our hypothesis is gesture memorization, our main dependent variable is the number of gestures that the user has to create and memorize. The two other dependent variables are colocalization (with, without = C1, C0) and affordance (with, without = A1, A0). We propose three levels of difficulty for number of gestures, with 1, 2 and 3 boxes to open for each condition. The design is a within-subjects user study with 2 repeated measures for: 3 (number of gestures) × 2 (affordance condition) × 2 (colocalization condition).

The columns in Figure 3 show the three levels of difficulty, i.e. 1, 2 and 3 gestures to memorize, for each condition (2G, 4G, and 6G) and the training step (1G). The rows show the two affordant conditions and the timeline of the experiment according to each step. Affordances are chosen to be visual clues on the boxes. For affordant boxes (A1), we have modeled six boxes with different visual clues: a hinge, a lock, covers with arrows or different types of edges (see Figure 3). Those are chosen in order to avoid similar opening gestures. For non affordant boxes (A0) the boxes are simply differentiated with colors. For colocalized manipulation (C1), the box is displayed in front of the user, on the virtual table at arms and hands reach. The user has to manipulate the box directly. For indirect manipulation (C0), the table and the boxes are located 1 meter far from the original position (see Figure 2). In this case the subject cannot reach the box and has to perform the gesture with the same posture and arm extent as for C1.

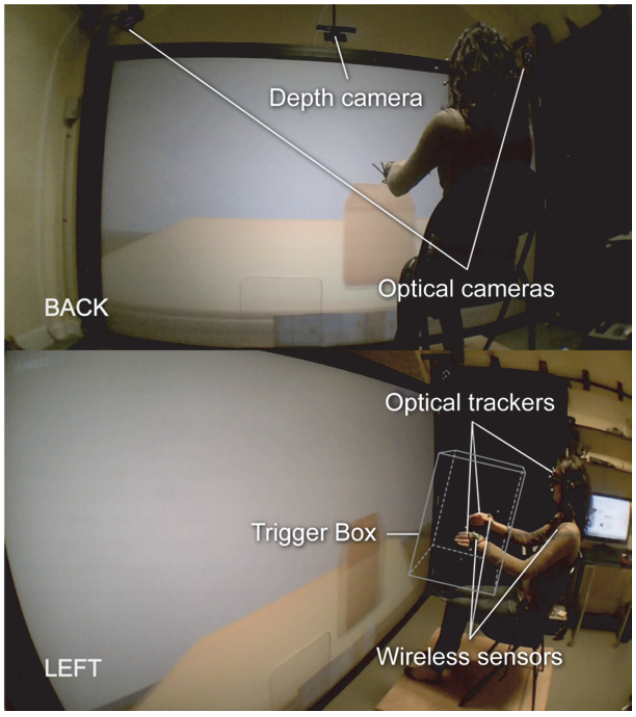

Figure1: Experimental setup of the user study.

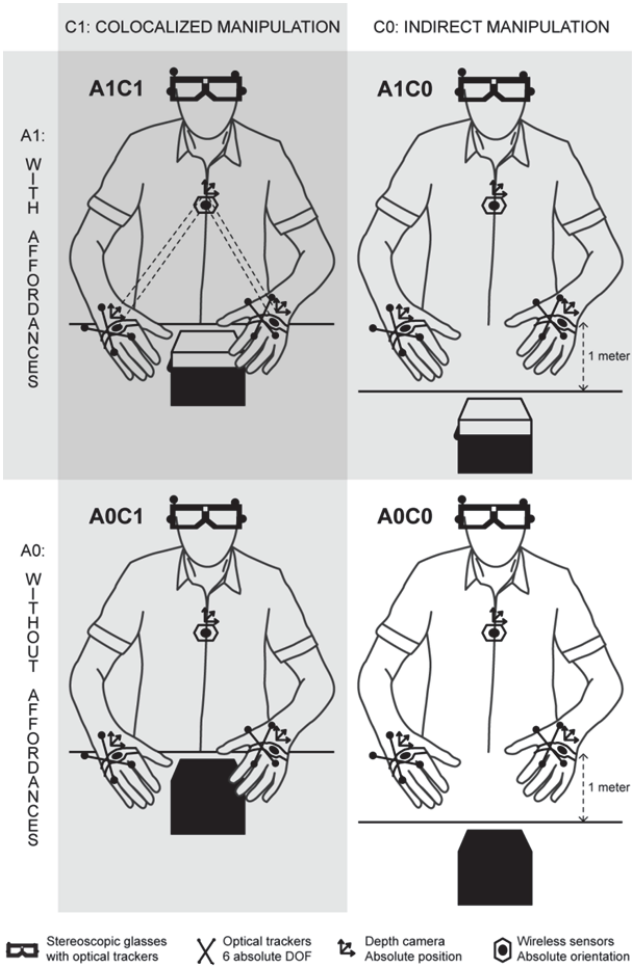

Figure 2: Boxes presented randomly in front of the user, with four combinations of conditions. A0: no visual clue on how the box opens only a specific color, A1: visual clue present. C0: distant manipulation, C1: direct manipulation.

4.2 Protocol

Prior to the experiment, the subjects were asked if they were left or right handed and invited to act as naturally as possible. The purpose of the experiment was not told to the subjects, neither the steps nor the number of boxes. Also, in order not to add bias to the evaluation of the memorization performance, we choose not to tell the subjects if they were correctly recalling gestures during the experiment.

During the **gesture creation** phase, a series of virtual boxes is displayed on a table in front of the user (see Figure 1). Subjects are asked to put their hands in the scene and propose a gesture to open each of the presented boxes. In order to stabilize the gesture, four repetitions of the created gesture are performed and only the last one is used for gesture learning. No indications are given whatsoever on the type of gestures to propose, neither clues on the differences between the boxes. N.B.: during this phase, boxes are presented only in direct manipulation condition.

During the **gesture interaction** phase, for each condition, the same set of boxes is presented randomly. The user is asked to perform the same gesture than the one he created previously according to the presented box.

Subjects are invited to put their hands on top of their thighs before and after performing each gesture. This allows relaxation time between actions, as recommended by Nielson et al. [27]. Gesture learning and following only occurs when they put their dominant hand within an invisible trigger volume surrounding the boxes so as to isolate the stroke part of the gesture (see Appendix). When the hand enters the trigger volume, a *start* sound is played and indicates that the motion capture begins. When the hand leaves the trigger volume, a *stop* sound indicates the end of the recording.

The three difficulty levels (number of gestures) are shown on the timeline in Figure 3. At the beginning of each step, during *gesture creation*, subjects have to propose a gesture four times (dotted white bars). Only the fourth one is recorded (solid white bars). Then, boxes are presented randomly during the *gesture interaction* phase. Gestures reproduced during the *gesture interaction* phase are recorded (black bars) and then compared to those proposed at the *gesture creation* phase. The experiment takes approximately 20 minutes per user to be completed. 2288 gestures were performed and 1430 were used in the analysis.

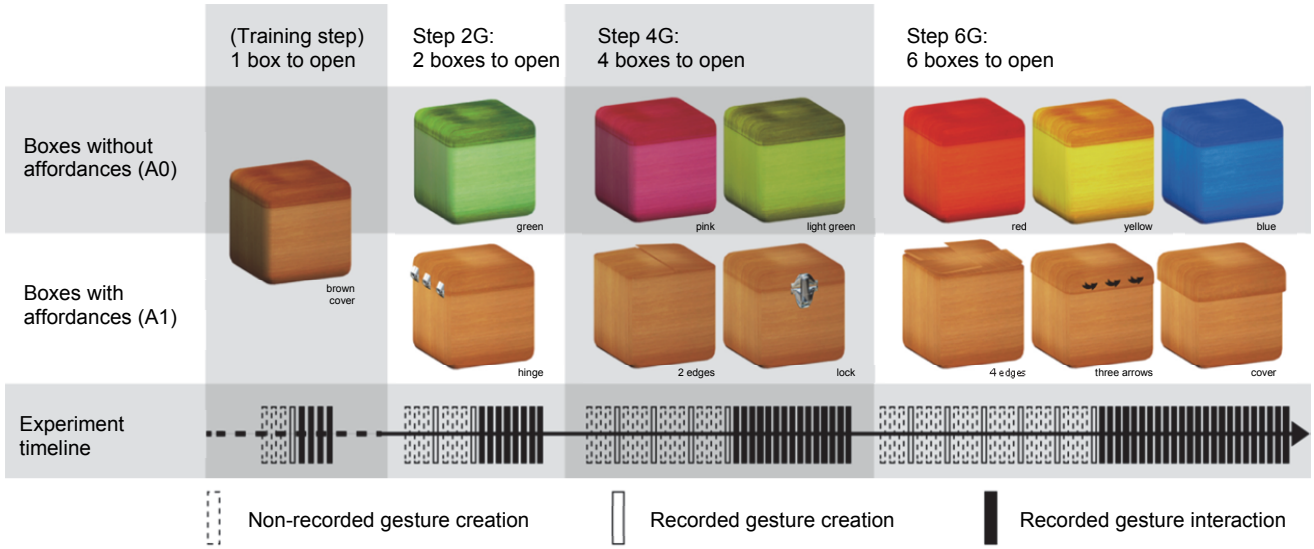

Figure 3: Set of boxes presented for each number of gesture steps and affordance conditions. The timeline shows the gesture production for each steps of the experiment.

4.3 Experimental setup

The experiment takes place in an immersive room with a back-projected wide screen (3.2m per 1.7m) with Full HD resolution (1920x1080@120Hz) and active stereoscopy. The head-tracking is performed with two optical infrared cameras (*Advanced Real-time Tracking GmbH, A.R.T.II* optical bodies and cameras). The setup is seen in Figure 1. To ensure accurate colocalisation, we have not used low-cost sensors for head-tracking, thought this would be an alternative. User's interpupillary distance is measured and set in the rendering engine, so the users perceive the VE with the proper proportions.

Absolute positions and orientations of trackers (6DOF) attached on each hand are recorded at 30Hz with the optical motion capture system, in order to get accurate information of gesture. Also, data from two low cost devices (*Microsoft Kinect camera* and *FAAST software* [28] and small gyroscopes *Movea MotionPods)* is recorded for the gesture recognition (see Appendix Section).

4.4 Subjects

Using this setup, we record data sets from 21 subjects (8 women and 13 men), aged from 21 to 34 years (M = 26.5) with interpupillary distances of 63.9 mm (56 mm to 70 mm). All subjects are tested for stereoscopic depth perception using the Wirt test, M = 82.3 % (10 % to 100 %). No subject had color-vision deficiency.

4.5 Measured parameters

Gesture comparison is not trivial. In order to adequately evaluate the performance, we compared gestures using three different criteria chosen during pre-tests:

- Reflection Time. It is the time from the moment the box is displayed to the end of completed *interaction gesture* minus the duration of the *created gesture*;
- Means of the Euclidean distances between *created gesture* and corresponding *interaction gesture* trajectories using the accurate motion capture system. All gestures are resampled to 200 points according to their time stamp (see graphs in Figure 7);
- The recognition rates outputted by a *Gesture Follower* Hidden Markov Model (HMM) and low-cost sensors when the gesture is completed (see Appendix).

5 Results and Discussion

Figures 4, 5 and 6 show the results for respectively reflection time, mean of Euclidian distances and recognition rates, without the training session (First step in Figure 3). Standard deviation is displayed as error bars. The means of the trials are analyzed with two-way repeated-measures ANOVAs.

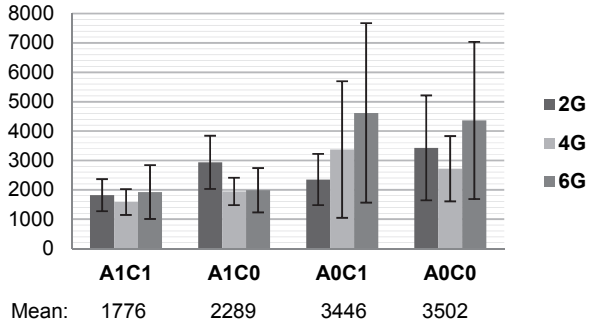

Figure 4: Reflection time in milliseconds. Lower is better.

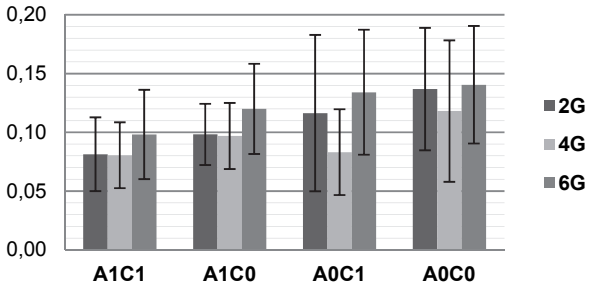

Figure 5: Mean of the Euclidean distances between gestures motion trajectories in meters (m). See plots in Figure 7 for gesture comparison. Lower is better.

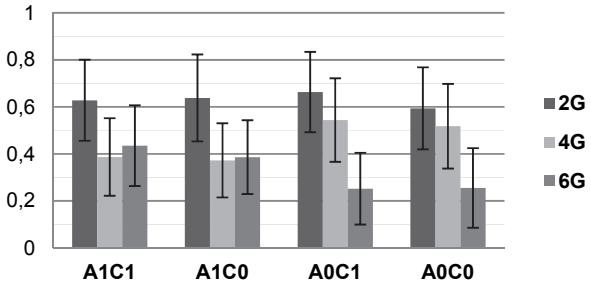

Figure 6: Gestures recognition rates (1 = 100 %). Higher is better.

5.1 Influence of Number of Gestures

The first trend that we observe is the decrease of user performance when the number of gestures increases. This trend can be seen for all non affordant boxes for the reflection time (A0, Figure 4) the precision (A0, Figure 5) and the recognition rates (A0, Figure 6).

A second observation is that we do not observe lower performances when affordance is present. The reflection time and the recognitions rates graphs show this same pattern on C1A1 and C0A1 columns (Figures 4 and 6). The means of gesture times with affordances are less than 2289 ms whereas, without affordances, means are superior to 3446 ms (see Mean in Figure 4). Without affordances, performance clearly lowers as the number of gestures increases (higher reflection time and lower recognition rates). It seems that the absence of affordance starts to impede memorization between 2G and 6G conditions that respectively include 1 and 3 non affordant boxes. Without affordances ANOVAs are non-significant between 2G and 4G, on reflection time $F(1,20) = .108$; $p = .746$ and on recognition rates $F(1,20) = .63$; $p = .437$. However there is a significant difference between 4G and 6G on time $F(1,20) = 8.89$; $p = .007$. The same trend is observed on recognition rate: they are twice as low (A0, Figure 6) and difference is significant $F(1,20) = 52.38$; $p < .001$. This suggests that affordance is needed to keep track of gestures starting from the 4G condition.

5.2 Influence of Affordance

It appears that affordances globally increase performances, whatever the number of gestures. They reduce reflection time regardless the number of gestures and the colocalization. ANOVA between A0 and A1 cases, independently of colocalization for all gestures sets, on reflection time is significant: $F(1,20) = 33.34$; $p < .001$. This is confirmed on the Euclidean distance too $F(1,20) = 4.88$; $p = .039$.

We can observe on plots A and C correct gesture replications with affordances. Plots E and F show a similar result, but the replication is less accurate, this is maybe due to fatigue at the end of the experiment (step 6G). Plots G and H clearly show two incorrect gesture replications: the user did not properly recall the original gesture.

5.3 Influence of Colocalization

The second trend is that performance seems independent of colocalization. For reflection time $F(1,20) = 2.22$; $p = .152$, however ANOVA on Euclidean distance shows a significant $F(1,20) = 25.74$; $p < .001$. This is resulting from an absolute offset in the mean Euclidian distance metric between the original gesture and its reproduction (see offset in plots B and D in Figure 7). In fact, we observed that during indirect manipulation, the users tend to reach towards the box, implying that the gesture is performed a bit farther than the original gesture.

5.4 Discussion

We have isolated affordance and colocalization as characteristics of schemata. Given the feeble role of colocalization on performance, schemata seem to be more related to visual affordances in a gestural memorization use case. We selected six different affordant ways to open boxes such as hinges, locks or covers with arrows and edges. One could argue that there are levels of affordances within these visual choices, which could have a positive or negative effect on the results, for visually small clues for example. This calls for a specific evaluation.

Also, during pre-tests we observed that some users tried to use mnemonics for non affordant boxes. In order to differentiate boxes, a digit was visually displayed on each box. Some users would base their gesture on the shape of the digit, literally drawing it in space or its first letter (example: an "f" for four). This made us choose colors as box differentiator in the actual tests to avoid too obvious mnemonics. Nevertheless, some users may have used mnemonics. We asked in a questionnaire about memorization strategies, six users reported associating the red box to a big red button or excitability, which produced for example slap or punch gestures. Somehow these users were trying to create their own affordances on the boxes. Choosing the right non-affordant appearance for boxes is an open question.

We could argue that the HMM recognition rates could only be algorithm related and would only depend on the number of different gestures for the system to recognize. However, we observe that the recognition rates are clearly stabilized for higher numbers of gestures when the affordance is present (Figure 6). This result shows that the origin of this better performance comes indeed from additional information to the users. Regarding the absolute recognition rates, we observed for some users that gesture recognition was difficult due to too small or co-articulated gestures. However, the mean recognition rates were satisfying despite the high variances. Also, the spatial offset of reproduced gestures observed between C0 and C1 during the experiment (as revealed with the Euclidean comparisons, see plots B and D in Figure 7) seems to have a limited effect on recognition rates.

Regarding the potential biases: the experiment boxes are presented randomly, however the time-line always follows the same sequence of 1, 2, 4 and 6 gestures. Even after the training session, for the 2G condition some subjects needed confirmation that the task follows the same pattern as training session when boxes were distant. This caused an increase of the measured reflection time as we can see on the first columns of indirect manipulation conditions (C0) in Figure 4.

We had to choose between mixing up A0 and A1 conditions and putting them into two separate series of tasks. We chose the first because our hypothesis is this would elicit more spontaneous behaviour: a series of non affordant boxes would have been understood as a classical memorization task.

Also, what is the role of the number of repetitions of the gesture in *gesture creation* phase? We asked the users about this in a subjective questionnaire. Nobody suggested that more gestures could help them to memorize.

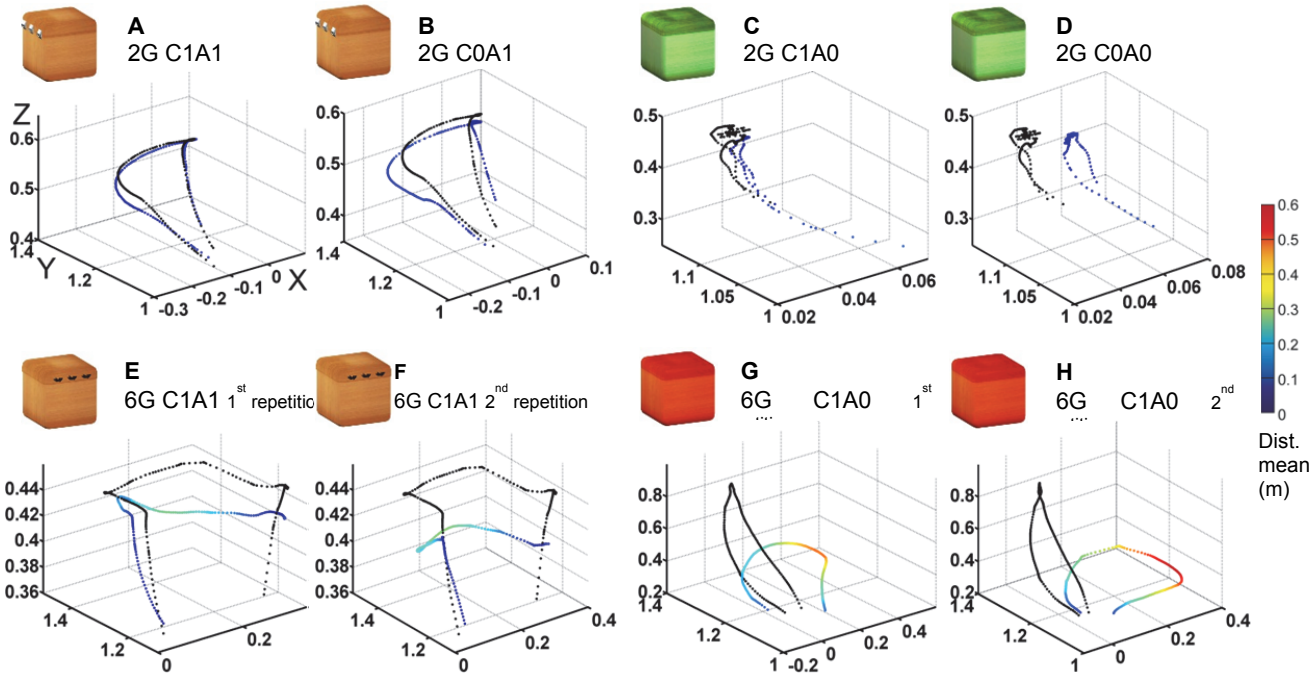

Figure 7: Comparison of gestures (room referential). The black curve is the original gesture. The colored curve codes the distance (blue to red) to the original curve.

6 Conclusion and Future Work

We have proposed an interactive process for users to create gestures for a truly dedicated interface. An implementation of the process using gesture following algorithms was proposed. We presented a study to explore the role of number of gestures, affordances and colocalization on gesture memorization and performance.

The results show a strong effect of affordances on gesture memorization but no apparent role of colocalization. It appears that remembering more than two imagined gestures isn't easy for one's working memory. In a context without affordances, the user is not able to recall more than two created gestures. However thanks to affordances users can rapidly recall three gestures. Thus, it would seem naturally possible to memorize more gestures, but how many? Everyday life situations suggest that gesture memorization could be unlimited since we perform a lot of gestures and learn/try new ones by referring to our innate abilities and personal experiences. Thus, the limit for an interactive application would be the ability of the gesture recognition system to adapt, both in number of gestures and variability of these gestures due to context or user.

As a consequence for rehabilitation applications, these results show that an indirect user interaction such as the ones that can be deployed at home would be relevant, using for instance low-cost tracking and TVs. Also, since colocalization has no effect on memorization, stereoscopy could not be necessary. This suggests head-tracking could be achieved with low-cost devices, though such hypothesis would require deeper analysis.

We did not evaluate the impact of visual, audio or haptic feedbacks which provide more information on the action performed. The natural continuation of this work is the study of the effect of visual animation while gesture is performed.

Indeed, in this experiment we chose not to provide a dynamic visual feedback since we wanted to focus on static visual affordances related to object appearance. Visual animation such as the box progressively opening with the gesture is another type of affordance that we can call dynamic. The gesture recognition system that we used allows a real-time following of the gesture progression so it is possible to do a second study with dynamic visual or audio affordances. It is the objective of our future work. Indeed, works from Varela [29] show that the process of learning is an *enactive* process: "cognitive structures emerge from the recurrent sensory-motor schemes which allow action to be guided by perception". In other words, to see is not to extract visual traits of the objects but to visually guide the action directed towards them.

Appendix: Gesture Analysis and Recognition

The objective of this appendix is to describe the HMM system that we choose for this particular study, which is *Gesture Follower* system (*add-on* for Max/*MSP)*[1]. Before this, we present a taxonomy of gestures, problematic of gesture recognition and existing solutions.

Many gestures classifications exist, the two studies [13], [14] described in section 2.1 use the same gesture taxonomy: the nature of gesture can be metaphorical, physical, symbolic or abstract. Those semantic meanings have to be properly identified and isolated during an action. The gesture phrase of this study are analyzed and cut regarding the work of Kendon on speech [30]. As described in Figure 8, the stroke is at the heart of the action. Preparation and retroaction phases can be either omitted, influenced or blend by the previous or next activities. This phenomenon is called coarticulation. In order to increase the gesture recognition rates, it seems appropriate to isolate correctly each different gesture and their stroke, especially during the training phase of machine learning.

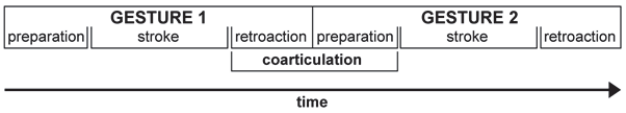

Figure 8: Decomposition of gesture based on formality and speech (adapted from Kendon [30]).

We look towards bringing gestural interaction via affordable VR systems at home or for small art companies using low-cost devices. Among other techniques, Hidden Markov Models (HMM) are known to best suit gesture recognition as shown in the survey [31]. Good recognition rates (> 80%) are also obtained with physically disabled people: Morrison and McKenna successfully recognize quavered gestures [32]. Schlömer et al. [33] also obtain good recognition results using acceleration from a Nintendo Wii Remote. However, most of these models output the probability of the recognized gesture after the gesture is completed.

Bevilacqua and his colleagues [34] proposed a HMM-based algorithm that continuously outputs parameters relative to the gesture, i.e. the time progression and the likelihood to the original gesture. The *Gesture Follower* can guess which gesture is currently performed or how close the current gesture is from the recorded ones. The typical usage is music conducting or score following, synchronizing physical gestures to sound files. Furthermore, this machine learning system allows training a gesture phrase with only one sample whereas most of the stochastic algorithms require multiple samples.

The *Gesture Follower* accepts normalized positions and orientation as descriptors of the hand movement. To do so, we decide to combine low-cost sensors that output these coordinates to perform the gesture recognition. Training and recognition of the HMM are done using 6 Degrees Of Freedom (DOF) of both hands. Pretests showed that placing each sensor on hands provides more DOFs than on wrists or forearms. It is also less intrusive than placing it on fingers. In order to capture hand movements independently from the user's upper body movements, the coordinates of each hand are torso referential (see dotted lines in Figure 2).

Positions and orientations of the hands and the torso are provided respectively by a *Microsoft Kinect camera* and FAAST software [28] and three small *Movea MotionPods* gyroscopes and a homemade VRPN server. One *MotionPod* is placed near the neckline and the two others are attached on the metacarpal of each hand (see Figure 2). About the precision and the noise the low-cost devices used, angles of the *MotionPod* sensors are quite precise (<0.01 degree, with very few drifting). The *Kinect camera* is also precise enough to perform the recognition (from 0.5 to 3 cm). However, in order to compensate the small noise and glitches, we apply filters such as a mean of the last five values combined to a threshold.

Internal benchmarks gave satisfying recognitions rates (> 70 %). The *Gesture Follower* accepts only one sample to be trained, but we observed variability in the first and second repetitions of gesture proposed by users. Indeed, those were sometimes used to tune up or refine the gesture. Thus, recording the forth gesture repetition was decided to record a more stable

[1] http://imtr.ircam.fr/imtr/Gesture_Follower

final gesture. The system has also been tested by a person with motor disability who was able to move his arms but not the fingers. Gestures were successfully recognized and performances measured were similar to those of the users group.

ACKNOWLEDGMENTS

The authors wish to thank all the subjects who participated in the pre-tests and the experiment.

REFERENCES

[1] M. E. Bobillier-Chaumon, S. Carvallo, F. Tarpin-Bernard, and J. Vacherand-Revel, "To adapt or standardize the human-computer interactions?," *Revue d'Interaction Homme-Machine*, vol. 6, no. 2, pp. 91–129, 2005.

[2] M. D. Good, J. a. Whiteside, D. R. Wixon, and S. J. Jones, "Building a user-derived interface," *Communications of the ACM*, vol. 27, no. 10, pp. 1032–1043, Oct. 1984.

[3] W. T. Freeman and C. D. Weissman, "Television control by hand gestures," in *IEEE Intl. Wkshp. on Automatic Face and Gesture Recognition*, 1995.

[4] D. A. Bowman, J. Chen, C. A. Wingrave, J. Lucas, A. Ray, N. F. Polys, Q. Li, Y. Haciahmetoglu, J. Kim, S. Kim, R. Boehringer, and T. Ni, "New Directions in 3D User Interfaces," *International Journal of Virtual Reality IJVR*, vol. 5, no. 2, pp. 3–14, 2006.

[5] M. Cabral, C. Morimoto, and M. Zuffo, "On the usability of gesture interfaces in virtual reality environments," in *Latin American conference on Human-computer interaction CLIHC*, 2005, pp. 100–108.

[6] A. J. D. Alon, S. Lazem, R. J. Beaton, D. Machaj, M. Schaefer, M. G. Silva, A. Leal, R. Hagan, and D. A. Bowman, "Evaluating natural interaction techniques in video games," in *IEEE 3D User Interfaces 3DUI*, 2010, pp. 11–14.

[7] D. A. Norman, "Natural user interfaces are not natural," *Interactions*, vol. 17, no. 3, pp. 6–10, 2010.

[8] J. J. LaViola and D. F. Keefe, "3D Spatial Interaction: Applications for Art, Design, and Science," in *ACM SIGGRAPH Courses*, Vancouver, British Columbia, Canada: ACM, 2011, p. 72.

[9] R. Oppermann, *Adaptive user support: ergonomic design of manually and automatically adaptable software*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1994.

[10] K. Laver, S. George, S. Thomas, J. Deutsch, and M. Crotty, "Virtual reality for stroke rehabilitation," *Cochrane Database of Systematic Reviews*, no. 9, 2011.

[11] S. Keates and P. Robinson, "The use of gestures in multimodal input," in *ACM conference on Assistive technologies ASSETS*, 1998, pp. 35–42.

[12] C. Vik, "Kinectar Performance Platform," 2011. [Online]. Available: http://ethnotekh.com/project/kinectar/.

[13] J. Ruiz, Y. Li, and E. Lank, "User-defined motion gestures for mobile interaction," in *Conference on Human Factors in Computing Systems SIGCHI*, 2011, pp. 197–206.

[14] J. O. Wobbrock, M. R. Morris, and A. D. Wilson, "User-defined gestures for surface computing," in *Conference on Human Factors in Computing Systems SIGCHI*, 2009, pp. 1083–1092.

[15] A. Baddeley, "Working memory: looking back and looking forward," *Nature Reviews Neuroscience*, vol. 4, no. 10, pp. 829–839, 2003.

[16] G. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological review*, vol. 63, no. 2, pp. 81–97, 1956.

[17] J. Farrington, "From the Research: Myths Worth Dispelling. Seven Plus or Minus Two," *Performance Improvement Quarterly*, vol. 23, no. 4, pp. 113–116, 2011.

[18] S. Wagner, H. Nusbaum, and S. Goldin-Meadow, "Probing the mental representation of gesture: Is handwaving spatial?," *Journal of Memory and Language*, vol. 50, no. 4, pp. 395–407, May 2004.

[19] J. Piaget, *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes.* University of Chicago Press, 1971.

[20] P. Fuchs, G. Moreau, and P. Guitton, *Virtual Reality: Concepts and Technologies*. 2011, p. 432.

[21] J. J. Gibson, *The ecological approach to visual perception*. Boston: Houghton Mifflin, 1979, p. 332.

[22] J. Mullaly, "IBM RealThings," in *Human factors in computing systems SIGCHI*, 1998, pp. 13–14.

[23] G. Rizzolatti, L. Fadiga, L. Fogassi, and V. Gallese, "The space around us," *Science*, vol. 277, no. 5323, pp. 190–191, 1997.

[24] G. Smets, P. Stappers, and K. Overbeeke, "Designing in virtual reality: implementing perception-action coupling with affordances," in *Virtual reality software and technology VRST*, 1994, pp. 97–110.

[25] M. R. Mine, P. B. J. Frederick, and C. H. Sequin, "Moving Objects In Space: Exploiting Proprioception In Virtual-Environment Interaction," in *Computer graphics and interactive techniques SIGGRAPH*, 1997, pp. 19–26.

[26] A. Paljic, S. Coquillart, J.-M. Burkhardt, and P. Richard, "A Study of Distance of Manipulation on the Responsive Workbench (tm)," in *Immersive Projection Technology Workshop IPT*, 2002.

[27] M. Nielsen, M. Störring, T. B. Moeslund, and E. Granum, "A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI," *Gesture-Based Communication in Human-Computer Interaction*, vol. 2915, pp. 105–106, 2004.

[28] E. A. Suma, B. Lange, S. Rizzo, D. M. Krum, and M. Bolas, "FAAST Flexible Action and Articulated Skeleton Toolkit," in *Virtual Reality Conference IEEE VR*, 2011, pp. 247 – 248.

[29] F. J. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: MIT Press, 1998, p. 338.

[30] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The relationship of verbal and nonverbal communication*, The Hague, The Netherlands: Mouton publishers, 1980, pp. 207–227.

[31] S. Mitra and T. Acharya, "Gesture Recognition : A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,*, vol. 37, no. 3, pp. 311–324, 2007.

[32] K. Morrison and S. McKenna, "Contact-Free Recognition of User-Defined Gestures as a Means of Computer Access for the Physically Disabled," in *Workshop on Universal Access and Assistive Technology*, 2002, pp. 99–103.

[33] T. Schlömer, B. Poppinga, N. Henze, and S. Boll, "Gesture recognition with a Wii controller," in *International conference on Tangible and Embedded Interaction TEI*, 2008, pp. 11–14.

[34] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, "Continuous realtime gesture following and recognition," *Gesture in Embodied Communication and Human-Computer Interaction*, vol. 5934, pp. 73–84, 2010.